智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

利用BERT和覆盖率机制改进的HiNT文本检索模型

邸剑, 刘骏华, 曹锦纲

引用本文:
邸剑, 刘骏华, 曹锦纲. 利用BERT和覆盖率机制改进的HiNT文本检索模型[J]. 智能系统学报, 2024, 19(3): 719-727.
DI Jian, LIU Junhua, CAO Jingang. An improved HiNT text retrieval model using BERT and coverage mechanism[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(3): 719-727.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202201020

DOI: 10.11992/tis.202201020
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20230926.1452.002

利用 BERT 和覆盖率机制改进的 HiNT 文本检索模型

邸剑[1,2], 刘骏华[1,2], 曹锦纲[1,2]
（1. 华北电力大学 控制与计算机工程学院, 河北 保定 071003; 2. 复杂能源系统智能计算教育部工程研究中心, 河北 保定 071003）

摘　要: 为有效提升文本语义检索的准确度, 本文针对当前文本检索模型衡量查询和文档的相关性时不能很好地解决文本歧义和一词多义等问题, 提出一种基于改进的分层神经匹配模型 (hierarchical neural matching model, HiNT)。该模型先对文档的各个段提取关键主题词, 然后用基于变换器的双向编码器（bidirectional encoder representations from transformers, BERT）模型将其编码为多个稠密的语义向量, 再利用引入覆盖率机制的局部匹配层进行处理, 使模型可以根据文档的局部段级别粒度和全局文档级别粒度进行相关性计算, 提高检索的准确率。本文提出的模型在 MS MARCO 和 webtext2019zh 数据集上与多个检索模型进行对比, 取得了最优结果, 验证了本文提出模型的有效性。

关键词: 基于变换器的双向编码器; 分层神经匹配模型; 覆盖率机制; 文本检索; 语义表示; 特征提取; 自然语言处理; 相似度; 多粒度
中图分类号: TP311　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)03−0719−09

中文引用格式: 邸剑, 刘骏华, 曹锦纲. 利用 BERT 和覆盖率机制改进的 HiNT 文本检索模型 [J]. 智能系统学报, 2024, 19(3): 719–727.
英文引用格式: DI Jian, LIU Junhua, CAO Jingang. An improved HiNT text retrieval model using BERT and coverage mechanism[J]. CAAI transactions on intelligent systems, 2024, 19(3): 719–727.

An improved HiNT text retrieval model using BERT and coverage mechanism

DI Jian[1,2], LIU Junhua[1,2], CAO Jingang[1,2]
(1. School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China; 2. Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding 071003, China)

Abstract: To effectively improve the accuracy of text semantic retrieval, an improved hierarchical neural matching model is proposed, which can solve the problems of text ambiguity and polysemy when using text retrieval models to measure the relevance of queries and documents. The model first extracts key subject words from each segment of the document and then encodes them into multiple dense semantic vectors using the BERT model. Afterward, the local matching layer introduced with the coverage mechanism is used for processing so that the model can calculate the correlation according to the local segment-level granularity and the global document-level granularity of the document and improve the retrieval accuracy. The proposed model is compared with multiple retrieval models on the MS MARCO and webtext2019zh datasets, and the optimal results obtained verify the effectiveness of the proposed model.

Keywords: bidirectional encoder representations from transformers; hierarchical neural matching model; coverage mechanism; text retrieval; semantic representation; feature extraction; natural language processing; similarity; multi-granularity

文本检索是信息检索最重要、最基础的方向, 它的功能是根据用户指定的查询从文档库中检索出相关的文档, 并根据相关度排序后返回给

收稿日期: 2022−01−13.　**网络出版日期:** 2023−09−27.
基金项目: 中央高校基本科研业务费专项 (2021MS085).
通信作者: 曹锦纲. E-mail: caojg168@126.com.

用户。文本检索中的核心问题是如何衡量查询和文档的相关性。对于大规模的文档库，查询表示的语义在许多文档的不同位置中都有不同程度的体现，其相关度需要综合文档的整体信息和局部信息来计算[1]。主流的处理方法是将查询和文档分别编码为稠密向量再进行复杂的相关性分数计算。由于用户输入的查询往往是高度提炼的简短语句，而文档中的信息复杂多样，文档中不同的位置其语义体现难以衡量，所以文档的单语义简单表示会造成语义丢失问题，导致检索结果不准确。

近年来，深度学习技术在许多领域取得了显著的成果，在文本检索领域，基于深度学习技术的文本检索模型明显优于基于统计的文本检索模型。在自然语言处理 (natural language processing, NLP) 领域，文本表示极其重要，在许多情况下，它对任务成败有决定性的影响。分层神经匹配模型 (hierarchical neural matching model, HiNT) 是一种较为典型的文本检索模型[2]，它采用了许多重要的设计思想，包括注意力机制、多粒度相关性等。但是语言本身是复杂的，一词多义现象十分普遍，原始的 HiNT 模型并没有完全解决这个问题，仍有提升空间。

为解决 HiNT 模型的不足，提升查询和文档的相关性计算准确度，本文对其不足进行改进。针对其计算量大、语言歧义处理不足和关键词语义表示不精确的缺点，本文利用 BERT(bidirectional encoder representation from transformers) 模型将查询关键词进行单个向量表示，对文档关键词进行多个向量表示，并引入覆盖率机制。在 MS MARCO 和 webtext2019zh 数据集上，改进的模型在各项关键指标上取得了良好的效果。

1　相关研究

1.1　重要检索模型

基于统计的概率检索模型是最先被广泛应用的。最经典的是 BM25 模型[3]，它利用 TF-IDF 计算索引关键词的权重，不仅融合了词频和文档长度因素，还有两个超参数用于调节模型，检索精度高，至今仍然是文本检索领域的基线算法[4]。但是该模型不能根据词义检索文档，因此在一些需要利用语义来检索内容的领域中应用效果不好。

基于深度学习的文本检索研究主要有两个方向：文本表示和文本交互。基于文本表示的模型的基本思想是将查询和文档表示成向量，利用其相似度 (通常是余弦相似度) 来计算相关度。基于文本交互的模型的基本思想是提取查询和文档的语义特征后进行排序学习，关注匹配信号、查询词和匹配多样化[5]。目前大部分模型都是基于这两者之一或两者不同程度的结合[6]。

微软提出的深度语义匹配模型[7](deep structured semantic models, DSSM) 是基于文本表示的典范，它将文本检索带入了全新的方向。其基本思路是将查询和文档都表示为同一语义空间的相同维度的语义向量，利用两者的余弦相似度计算相关度，最终返回有序的文档集合。此模型及其衍生模型在短文本检索方面超越了传统的 BM25 模型。

DSSM 由 3 层组成，分别是输入层、表示层和匹配层。输入层可视为对文本信息的预处理，将原始的文本转换为独热编码后使用哈希编码 (word hashing) 压缩降维，产生便于后续处理的较为稠密的语义向量。表示层将输入层产生的预处理向量映射至同一语义空间，分别生成查询和文档的稠密语义向量。匹配层较为简单，它将表示层产生的两个语义向量进行余弦相似度计算，计算结果即查询和文档的相似度分值。其中表示层是模型最核心的部分，它直接影响查询和文档的相关度分值，对文本检索结果的准确程度影响巨大。由于 DSSM 采用词袋模型表示文本，不能很好地利用文本上下文，人们针对此问题提出了多种多样的变种模型，如 CNN-DSSM[8] 和 LSTM-DSSM[9] 等。

Guo 等[10] 提出的 DRMM 模型 (deep relevance matching model) 是基于查询和文档交互的代表模型，其基本思想是利用查询和文档构造匹配矩阵，提取其特征后再结合从 term gating 得到的关键词权重进行加权计算，最终输出相关度分值。此模型及其衍生模型在长文本方面超越了传统的 BM25 模型。

DRMM 利用查询和文档的词级别向量，词向量由 word2vec 生成，首先将查询和文档的词向量通过余弦相似度函数构建相似度交互矩阵，其中矩阵维度和文档的长度相等。然后通过直方图函数将不规整的交互矩阵转换为规则的直方图，即将局部的交互矩阵映射为全局的直方图。最后提取直方图的特征，通过 term gating 整合词的权值来计算查询和文档的相似度分值。原始的 DRMM 模型忽视了文档上下文信息对匹配的影响，为此人们提出了优化的变体模型，如 KNRM、PACRR[11]、Co-PACRR[12] 等。

1.2　文本表示

深度学习技术广泛应用于 NLP 领域，文本表

示是各种复杂任务的关键步骤, 它指的是利用适当的表示方法将文本转化成可量化的数学表示[1]。通常根据文本表示的不同粒度可将文本表示为文档级别向量、段级别向量和词级别向量等, 本文针对的是词级别向量。由于词向量粒度小, 既可以方便地利用文档上下文信息, 也适合构造更高层次的表示, 所以基于词向量的模型在 NLP 各项领域中大放异彩。

分布式词向量曾经掀起了文本表示的热潮, 词的语义可以用其上下文表示为稠密的向量, 词与词的相似程度即向量距离[13]。典型的是 word2-vec 模型, 它常用连续词袋(continuous bag of words, CBOW)模型和 skip-gram 实现, 缺乏对文档整体的学习[14]。不久, Glove 算法被提出, 它结合了文本整体和上下文的语义信息。但是分布式词向量生成的是静态不变的, 不能解决一词多义和 OOV 问题, 也不能对下游任务做专门的优化。针对此问题, 基于上下文的 BERT 模型大行其道。BERT 是基于双向的端对端的 Transformer 模型[15], 可以利用双向的上下文和位置信息, 也可以解决文本序列长距离依赖问题。它经过预训练, 生成通用的模型参数, 之后针对特定的任务进行微调, 这样既简化训练流程, 也提高了模型泛化能力[16]。

1.3　原始 HiNT 模型

HiNT 是一种层次化的深度神经网络模型, 将文档全局信息和文档段信息通过竞争机制很好地结合起来, 可以解决长文档和短文档之间竞争存在潜在偏向的问题。它由两部分组成, 分别是局部匹配层和全局决策层。局部匹配层通过查询与文档各段之间的语义匹配来产生文档的局部相关信号。全局决策层将局部信号聚合成不同的粒度, 通过相互竞争以决定最终的相关性分值。总体结构如图 1 所示。

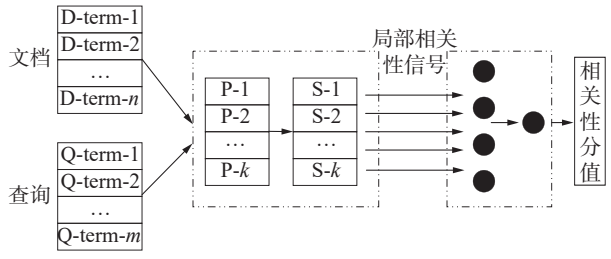


图 1　HiNT 模型总体结构

Fig. 1　Overall structure of the original HiNT model

1.3.1　局部匹配层

局部匹配层通过计算查询与文档中每个段之间的相关性匹配, 生成一组段级相关信号。每个文档 $\boldsymbol{D}$ 首先表示为多个段 P_i, $\boldsymbol{D}=[P_1 \ \ P_2 \ \ \cdots \ \ P_K]$, 其中 K 表示文档中的段总数。然后, 段级相关性信号 $\boldsymbol{E}=[e_1 \ \ e_2 \ \ \cdots \ \ e_K]$ 由匹配函数 f 根据查询 $\boldsymbol{Q}$ 和对应每段的相似度分值产生, 如下式所示:

$$e_i = f(P_i, \boldsymbol{Q}), \quad i = 1,2,\cdots,K \tag{1}$$

P_i 的表示和匹配函数 f 的设定是重点问题。HiNT 使用固定大小的窗口来表示文档, 利用双向空间 GRU(gate recurrent unit)进行匹配。局部匹配层工作原理如图 2 所示[2]。

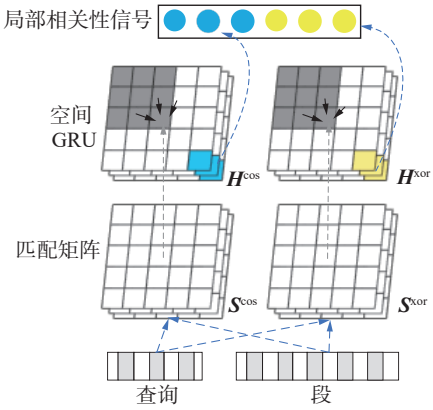


图 2　局部匹配层工作原理

Fig. 2　Local matching layer work schematic

局部匹配层工作原理形式化描述为: 对于给定的查询 $\boldsymbol{Q}$ 和文档 $\boldsymbol{D}$, $\boldsymbol{Q}$ 由 M 个关键词组成 $\boldsymbol{Q}=\left[\boldsymbol{w}_1^{(Q)} \ \ \boldsymbol{w}_2^{(Q)} \ \ \cdots \ \ \boldsymbol{w}_M^{(Q)}\right]$, $\boldsymbol{D}$ 由 N 个关键词组成 $\boldsymbol{D}=\left[\boldsymbol{w}_1^{(D)} \ \ \boldsymbol{w}_2^{(D)} \ \ \cdots \ \ \boldsymbol{w}_N^{(D)}\right]$。同时 $\boldsymbol{D}$ 由多个段 $\boldsymbol{P}$ 组成, 而每个段以固定大小 L 分割, 即 $\boldsymbol{P}=\left[\boldsymbol{w}_1^{(P)} \ \ \boldsymbol{w}_2^{(P)} \ \ \cdots \ \ \boldsymbol{w}_L^{(P)}\right]$。词之间的相似度计算使用 cosine 余弦相似度计算法和 xor 计算法结合表示, 分别生成两个相似度矩阵 $\boldsymbol{M}_{ij}^{\cos}$ 和 $\boldsymbol{M}_{ij}^{\text{xor}}$。两个矩阵的关键思想是明确地区分语义匹配信号和精确匹配信号, 分别如下式所示:

$$\boldsymbol{M}_{ij}^{\cos} = \frac{\boldsymbol{w}_i^{(Q)} \boldsymbol{w}_j^{(P)}}{\left|\boldsymbol{w}_i^{(Q)}\right| \cdot \left|\boldsymbol{w}_j^{(P)}\right|} \tag{2}$$

$$\boldsymbol{M}_{ij}^{\text{xor}} = \begin{cases} 1, & \boldsymbol{w}_i^{(Q)} = \boldsymbol{w}_j^{(P)} \\ 0, & \text{其他} \end{cases} \tag{3}$$

将 $\boldsymbol{M}_i$ 的每个元素扩展成三维向量 $\boldsymbol{S}_{ij}=[\boldsymbol{x}_i \ \ \boldsymbol{y}_j \ \ \boldsymbol{M}_{ij}]$, $\boldsymbol{x}_i$、$\boldsymbol{y}_j$ 分别是 $\boldsymbol{w}_i^{(Q)}$ 和 $\boldsymbol{w}_j^{(P)}$ 经过共享的转换矩阵 $\boldsymbol{W}_s$ 得出:

$$\boldsymbol{x}_i = \boldsymbol{w}_i^{(Q)} * \boldsymbol{W}_s \tag{4}$$

$$\boldsymbol{y}_j = \boldsymbol{w}_j^{(P)} * \boldsymbol{W}_s \tag{5}$$

之后将这两个匹配矩阵通过 spatial GRU 处理, spatial GRU 是融合了空间信息的 2 维 GRU,

$\vec{\boldsymbol{H}}_{ij}$由左、上、左上 3 个隐层的状态和当前匹配分数$\boldsymbol{S}_{ij}$组成，右箭头表示左、上、左上的处理方向：

$$\vec{\boldsymbol{H}}_{ij}^{\cos}=g\left(\vec{\boldsymbol{H}}_{i-1,j}^{\cos},\vec{\boldsymbol{H}}_{i,j-1}^{\cos},\vec{\boldsymbol{H}}_{i-1,j-1}^{\cos},\boldsymbol{S}_{ij}^{\cos}\right) \tag{6}$$

$$\vec{\boldsymbol{H}}_{ij}^{\mathrm{xor}}=g\left(\vec{\boldsymbol{H}}_{i-1,j}^{\mathrm{xor}},\vec{\boldsymbol{H}}_{i,j-1}^{\mathrm{xor}},\vec{\boldsymbol{H}}_{i-1,j-1}^{\mathrm{xor}},\boldsymbol{S}_{ij}^{\mathrm{xor}}\right) \tag{7}$$

最后一个隐层的状态$\vec{\boldsymbol{H}}_{M,L}^{\cos}$和$\vec{\boldsymbol{H}}_{M,L}^{\mathrm{xor}}$为两个相似度矩阵经过 spatial GRU 处理后的输出，将它们和经逆方向的 spatial GRU 处理的输出$\overleftarrow{\boldsymbol{H}}_{M,L}^{\cos}$和$\overleftarrow{\boldsymbol{H}}_{M,L}^{\mathrm{xor}}$拼接作为局部匹配层的最终输出$e$，左箭头表示右、下、右下的处理方向：

$$e=\left[\left[\vec{\boldsymbol{H}}_{M,L}^{\cos},\vec{\boldsymbol{H}}_{M,L}^{\mathrm{xor}}\right],\left[\overleftarrow{\boldsymbol{H}}_{M,L}^{\cos},\overleftarrow{\boldsymbol{H}}_{M,L}^{\mathrm{xor}}\right]\right] \tag{8}$$

1.3.2　全局决策层

全局决策层将信号累积成不同的粒度，它们相互竞争以计算最终的相关性。

HiNT 的全局决策层使用混合网络结构，将局部匹配层的K个向量接入非线性映射层 tanh 得到$\boldsymbol{v}_t$：

$$v_t=\tanh\left(\boldsymbol{W}_v e_t+\boldsymbol{b}_v\right) \tag{9}$$

同时输入 Bi-LSTM 层来捕捉这些向量之间的关系得到h_t。用 K-max pool 方法处理v_t和h_t，取前k个最大值拼接在一起后输入多层感知机（multilayer perceptron, MLP）中，得到最终的相似度分值。

2　改进的 HiNT 模型

改进的 HiNT 模型由 4 部分组成，分别是段关键字提取、BERT 多向量表示、结合覆盖率机制的局部匹配层和改进的全局决策层。改进的 HiNT 模型总体结构如图 3 所示。$\boldsymbol{H}^{\cos}$ 和 $\boldsymbol{H}^{\mathrm{xor}}$ 分别表示空间 GRU 在 $\boldsymbol{S}^{\cos}$ 和 $\boldsymbol{S}^{\mathrm{xor}}$ 上隐层的状态，$\boldsymbol{S}^{\cos}$ 和 $\boldsymbol{S}^{\mathrm{xor}}$ 分别表示将语义匹配矩阵 $\boldsymbol{M}^{\cos}$ 和精确匹配矩阵 $\boldsymbol{M}^{\mathrm{xor}}$ 的每个元素扩展为三维向量后的结果。

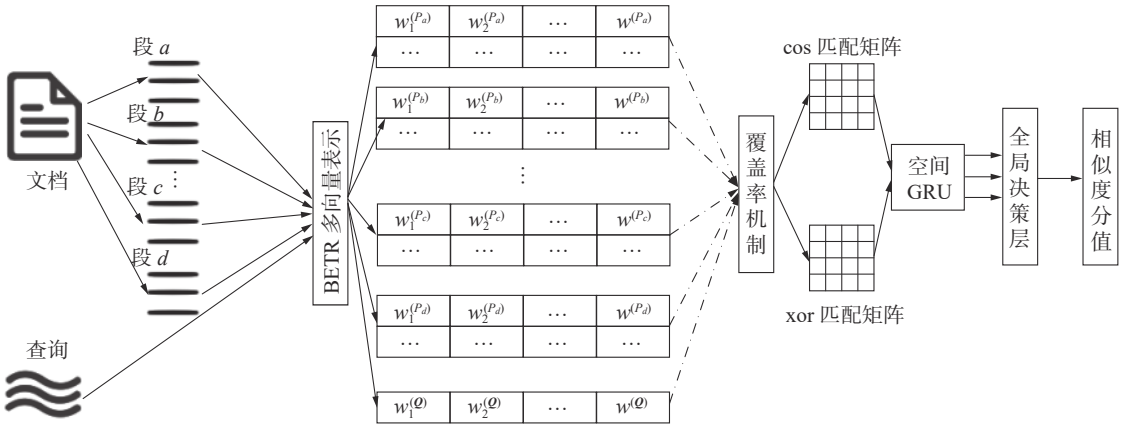


图 3　改进的 HiNT 模型总体结构

Fig. 3　Overall structure of the improved HiNT model

2.1　段主题关键字提取

HiNT 模型将段的所有词都进行计算，不仅计算量很大，而且融入很多与主题偏离的语义信息，不利于检索。改进后的模型只根据文档中每个段的主题关键字进行分析，极大减少了计算量。现如今，主题关键词提取技术十分成熟，本文使用 TextRank 方法，由 jieba 实现。

2.2　BERT 多向量表示

BERT 模型是预训练的语言表征模型，使用遮挡语言模型对双向 Transformer 进行预训练来生成双向语言表征[17]。本文的 BERT 语义表示模块由 3 层组成，分别是输入层、编码层和输出层，查询和文档的关键词共享这些参数。

输入层将提取的关键词根据词表得到一维词嵌入，根据词的位置生成相应的段嵌入和位置嵌入并输入模型，如图 4 所示。

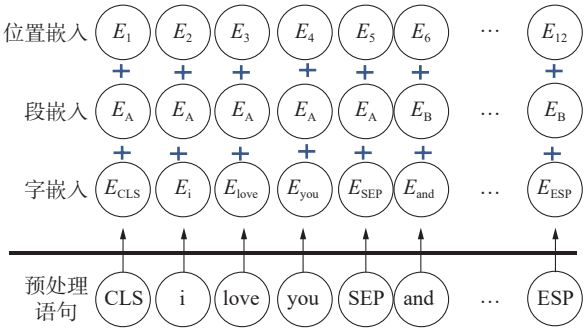


图 4　BERT 输入层

Fig. 4　Input layer of BERT

编码层利用自注意力机制处理输入层生成的词嵌入，将采用多头注意力机制和残差连接机制结构的多层 Transformer 单元堆叠，然后进行归一化。将输入转化成向量表示，最后聚合为前馈神经网络的输入[15]。

如图 5 所示，自注意力机制层将 BERT 输入

的序列X分别乘以 3 个权重向量$\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$和$\boldsymbol{W}^{v}$, 将其转换为不同的向量表示, Query($\boldsymbol{Q}$)、Key($\boldsymbol{K}$) 和 Value($\boldsymbol{v}$) [18]:

$$\begin{cases} \boldsymbol{Q} = X \times \boldsymbol{W}^{Q} \\ \boldsymbol{K} = X \times \boldsymbol{W}^{K} \\ \boldsymbol{v} = X \times \boldsymbol{W}^{v} \end{cases} \tag{10}$$

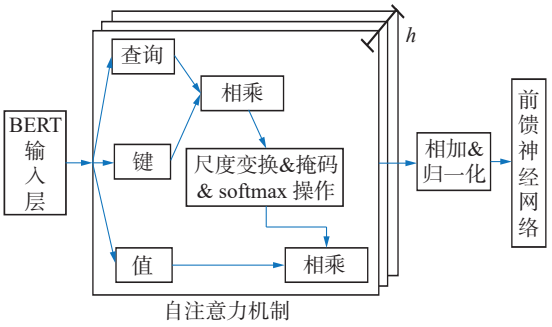


图 5　BERT 编码层

Fig. 5　Encoding layer of BERT

之后进行特征提取, 得到其矩阵表示, 其中缩放因子d_k用于控制映射范围, 特征提取方法为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{v}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)v \tag{11}$$

输出层将多头注意力机制的每个 Attention 权重信息进行整合, 使生成的词向量尽可能地结合上下文语义信息, 有利于后续处理。经过 BERT 网络后, 产生查询词嵌入表示$\left\{\boldsymbol{w}_i^{(Q)}\right\}_{i=1}^{n} \in \mathrm{R}^{n\times h}$和文档词嵌入表示$\left\{\boldsymbol{w}_i^{(D)}\right\}_{i=1}^{m} \in \mathrm{R}^{m\times h}$, h是隐层维度。

2.3　结合覆盖率机制的局部匹配层

大多数情况下, 查询的关键词长度短, 语义凝练, 因此 BERT 模型输出的 [CLS] 对应的编码可作为查询的向量表示。但是当文档的文本长度较长, 语义较为复杂, 主题多样时, 将文档关键词表示为单个向量表示难以融合其语义[19]。为此将文档关键词表示为多个有足够差异性的向量表示, 计算方法为

$$e_d^j = \sum_{i=1}^{m} \omega_i^j \cdot \boldsymbol{w}_i^{(D)}, \qquad j \in |1,z| \tag{12}$$

$$\omega_1^j, \omega_2^j, \cdots, \omega_m^j = \text{softmax}\left(\boldsymbol{w}_1^{(D)} \cdot c^j, \boldsymbol{w}_2^{(D)} \cdot c^j, \cdots, \boldsymbol{w}_m^{(D)} \cdot c^j\right) \tag{13}$$

式中: c^j是第j个注意力层的参数; z是超参数, 代表文档的语义向量表示个数。为保证这些向量覆盖文档的全局而非局部, 本文引入了覆盖率机制 (coverage), 该机制在文本摘要领域取得了优异成绩, 许多研究者将其扩展到了文本检索领域[20-22], 并取得了一定效果。记录历史j个注意力权重分布的累计, 并更新注意力权重分布[20], 以防止模型再次将注意力集中于原位置:

$$\omega_i^{j+1} = \omega_i^{j+1} - \sum_{l=1}^{j} \omega_i^l \tag{14}$$

这样, 局部匹配层的相似度矩阵就从一个变为多个, 将这些矩阵按照元素位置去掉最高值和最低值后计算平均值进行叠加, 生成更好地整合上下文信息的相似度矩阵, 再经过 spatial GRU 处理。

2.4　改进的全局决策层

HiNT 模型将向量经过 Bi-LSTM(long short-term memory)进行处理, 本文改为 Bi-GRU 模型。GRU 比 LSTM 的参数量少、结构简单, 只有 2 个门控单元, 可以减少过拟合的风险并且计算效率高, 因此它的可扩展性有利于构建较大的模型[23]。

损失函数设为 pairwise ranking loss, 即三元组(q,d^+,d^-), 检索分数高的d^+相对于查询的排名高于检索分数低的d^-, 具体计算公式为

$$\mathcal{L}(q,d^+,d^-;\theta) = \max(0, 1 - s(q,d^+) + s(q,d^-)) \tag{15}$$

式中: $s(q,d)$是(q,d)的相关性得分, θ为局部匹配层和全局决策层的参数。

3　实验和分析

为了验证提出模型的有效性, 本文在 2 个数据集上与多个模型进行了对比实验。

3.1　实验设置

本文的实验环境配置如表 1 所示。

表 1　实验环境配置

Table 1　Experimental environment configuration table

名称	详细信息
CPU	Intel(R) Core(TM) i7-10700 CPU @ 2.90 GHz
GPU	NVIDIA GeForce RTX 3070Ti/16GB
操作系统	Ubuntu 16.04
软件	Python3.6、PyTorch、TextNet

本文使用 MS MARCO[24] 和 webtext2019zh 数据集进行对比实验。MS MARCO 是 8.8×10^6 个网页的集合, 每个查询都对应于部分不同程度相关的文档。MS MARCO 数据集包含 300 万左右的文档, 训练集含有将近 40 万个查询, 十分适合文本检索任务。webtext2019zh 是大规模高质量的社会问答数据集, 由于在深度学习领域, 文本匹配和智能问答具有诸多相似性, 所以该数据集适用于本文的实验。

模型参数方面, 使用 Adam 优化算法, 学习率为 0.001, 段大小为 100, 向量表示个数z设为 4, 采用 5 折交叉验证法进行验证。

3.2　评价指标

本文使用召回率 Recall 和平均倒数排名 (mean reciprocal rank, MRR) 作为评价指标。

Recall: 样本中的正例被正确预测的比率, 本文取前 50 和 1 000 的检索结果计算。

MRR: 将检索结果中正确检索位置排名的倒数作为评价质量, 其中, M_n表示平均倒数排名, Q是总查询数, r_i是最佳查询位置:

$$M_n = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{r_i} \tag{16}$$

3.3　对比模型

与本文改进模型比较的模型如下:

1) BM25、DSSM、RMM、HiNT 模型上文已有介绍。

2) Co-PACRR[12]: 考虑了文本的位置信息, 它加入一个卷积层来提取位置信息, 卷积网络可以更有效地提取词与词之间的特征, 并将局部上下文和全局信息通过串联模式生成文本语义向量表示。

3) DeepCT[25]: 采用深度上下文词语权重框架, 利用 BERT 的文本表示映射到句子和段落的上下文的词权重, 赋予不用文本中的相同词以不同的权重, 将其与 BM25 结合。

4) doc2query[26]: 基于文本扩展技术的稀疏检索算法, 针对每个文档都会做简单预测以关联相关查询, 本质上是一个 Seq2Seq 生成模型, 检索速度很快, 一定程度上兼具稀疏检索的速度和稠密检索的质量。

5) RepBert[27]: 采用基于双塔结构的 BERT 作为编码器来分别构建查询和文档的语义表示, 在此之上进行相似度计算。

3.4　结果和分析

表 2、3 分别给出了各模型在 MS MARCO 数据集和 webtext2019zh 数据集的测试结果。

表 2　各个模型在 MS MARCO 数据集上的结果

Table 2　Results of each model on the MS MARCO dataset

模型	MRR@10	Recall@50	Recall@1k
BM25	0.187	0.592	0.857
DSSM	0.177	0.441	0.883
DRMM	0.211	0.595	0.904
Co-PACRR	0.244	—	0.913
doc2query	0.215	0.644	0.891
DeepCT	0.243	0.690	0.910
RepBert	0.304	—	0.943
HiNT	0.301	0.744	0.931
本文模型	**0.318**	**0.752**	**0.944**

表 3　各个模型在 webtext2019zh 数据集上的结果

Table 3　Results of each model on the webtext2019zh dataset

模型	MRR@10	Recall@50	Recall@1k
BM25	0.155	0.520	0.784
DSSM	0.160	0.525	0.788
DRMM	0.153	0.504	0.791
Co-PACRR	0.155	—	0.803
doc2query	0.150	0.517	0.813
DeepCT	0.188	0.582	0.821
RepBert	0.199	—	0.855
HiNT	0.206	0.639	**0.883**
本文模型	**0.211**	**0.642**	0.881

从表 2 可以看出, 除了 DSSM 外, 其他模型都比传统的 BM25 方法性能强, 这表明仅利用简单的余弦相似度衡量查询和文档的相关度是片面的, 而不能融合上下文信息的 DRMM 模型也不能取得很好的性能; HiNT 的分层机制比 Co-PACRR 的卷积机制的 MRR@10 和 Recall@1k 分别提升了 5.7% 和 1.8%, 说明它更能准确地提取局部和全局特征; doc2query 和 Co-PACRR 与本文提出模型存在一定差距, 因为词扩展机制局限性很大, 难以消除歧义, 所以性能并不好; 本文提出的模型在 MS MARCO 数据集上比原始 HiNT 在 MRR@10、Recall@50 和 Recall@1k 指标上分别提升了 1.7%、0.8% 和 1.3%, 且和基于 BERT 的 RepBert 模型性能相比 MRR@10 提升了 1.4%, 说明本文提出的覆盖率机制改进的局部匹配和全局决策机制是有效的, 使用 BERT 产生的词向量缓解了 RepBert 将整个文档转换成一个语义向量表示会造成语义信息丢失的问题, 所以性能有显著提升。

从表 3 可以看出, 本文提出模型相较于 BM25, 在 MRR@10、Recall@50 和 Recall@1k 评价指标上分别提升了 5.6%、12.2%、9.7%, 同时可以发现使用 BERT 的 DeepCT、RepBert、HiNT 和本文提出的模型比不使用 BERT 的其他模型在各个指标上提升明显, 说明 BERT 机制可增强文本语义表示进而提升检索精度。在 webtext2019zh 数据集上, 本文提出的模型比原始 HiNT 在 MRR@10 和 Recall@50 指标上分别提升了 0.5%、0.3%, 在 Recall@1k 指标上略低于 HiNT。总的来说, 本文提出的方法较其他各种方法效果更好。在 webtext2019zh 数据集上各对比模型相对于 BM25 基线的指标提升不如在 MS MARCO 数据集上的明显, 这是因为该数据集所涉范围庞杂, 文本较为口语化, 相比于规范的 MS MARCO 数据集难以提取其语义

特征。

3.5　消融实验

为了验证提出模型各部分的有效性, 本文在 MS MARCO 数据集上进行了消融实验。消融实验对比模型包括 HiNT、仅使用段主题关键词提取改进 (+top)、仅使用 BERT 改进 (+bert)、仅使用覆盖率机制改进 (+coverage)、仅用 Bi-GRU 改造全局决策层 (+bigru) 等方法。其中+top 指的是匹配矩阵根据段落的主题关键词计算, 其余部分不变; +bert 指的是仅用 BERT 作为词语义表示, 而不使用覆盖率机制改动局部匹配层, 也不提取段主题词; +coverage 指的是使用原始方法的词向量输入, 但是局部匹配层会结合覆盖率机制, 不提取段主题词。消融实验结果如表 4 所示。

表 4　在 MS MARCO 数据集上的消融实验结果

Table 4　Results of ablation experiments on MS MARCO dataset

模型名称	MRR@10	Recall@50	Recall@1 000
HiNT	0.301	0.744	0.931
+top	0.305	0.751	0.927
+bert	0.301	0.747	0.911
+coverage	0.315	0.748	**0.945**
+bigru	0.299	0.737	0.933
本文模型	**0.318**	**0.752**	0.944

从表 4 可以看出, +bert 在 Recall@50 比 HiNT 模型提升了 0.3%, 但在 Recall@1k 上下降了 2%, +coverage 在 MRR@10、Recall@50 和 Recall@1 000 比 HiNT 模型分别提升了 1.4%、0. 4% 和 1.4%, 说明覆盖率机制比 BERT 语义表示更能对模型产生积极影响, 这可能是由于数据集的文档较为散乱导致的, 其文档篇幅或大或小, 文档的书写质量良莠不齐, 很多文档的语义偏离中心语义导致的。加入覆盖率机制后, 模型性能大幅提高, 因为它可以极大程度解决语义偏离问题。仅提取段主题关键词和 Bi-GRU 改造全局决策层, 虽然在各指标上没有明显提高, 但是可以减少模型的计算量。本文提出模型在 MRR@10 和 Recall@50 达到了最优, 而 Recall@1k 指标上与+coverage 相当。可见模型性能提升是各部分共同作用的结果, 说明了改进的有效性。

本文还对超参数z对模型性能的影响进行了分析。根据式 (12) 和 (14) 可知超参数z会影响历史 j 个注意力权重分布的累计, 进而影响检索精度。图 6 是改进后的 HiNT 模型在z值分别为 2、4、6 和 8 条件下的性能结果图, 可以看出模型性能随z值先增大后减小。因为z值过小词向量不能产生适当的语义差距, z值过大时语义会分散在多个向量表示中。即累计注意力权重与文本语义的主题数量有一定联系, 当文本语义简单, 主题少时z应该较小; 当文本语义复杂, 主题多时z应该较大。否则构造匹配矩阵时会出现语义相似度下降的情况, 所有z应该适中。

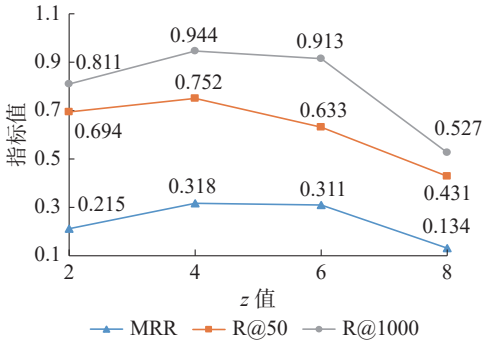


图 6　各指标在 MS MARCO 数据集上随z值的变化

Fig. 6　Variation of each indicator with z value on MS MARCO dataset

4　结束语

本文提出了一种改进的 HiNT 模型, 此模型在已有的模型基础上参照文本表示和文本交互匹配的思想, 进行段关键词提取后, 利用 BERT 的多向量表示进行关键词表示, 结合覆盖率机制, 比原始 HiNT 模型更好地融合局部上下文和文档整体信息, 相似度计算更为准确, 对检索任务具有重要意义。实验结果表明, 相比文本检索的典型模型, 本文提出的方法的各项指标有显著提升。但是该方法只能从已有的文档库中学习, 不具备拓展能力, 不能对新内容进行语义检索, 可以引入知识图谱来解决此问题。另外, 针对长短文本的检索任务也需要分别考虑, 以做进一步研究。

参考文献:

[1]　STADIG I, SVANBERG T. Overview of information retrieval in a hospital-based health technology assessment center in a Swedish Region[J]. International journal of technology assessment in health care, 2021, 37(1): e52.

[2]　FAN Yixing, GUO Jiafeng, LAN Yanyan, et al. Modeling diverse relevance patterns in ad-hoc retrieval[C]// SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor: ACM, 2018: 375–384.

[3]　ROBERTSON S E, JONES K S. Relevance weighting of

search terms[J]. Journal of the American society for information science, 1976, 27(3): 129–146.

[4] ROBERTSON S E, WALKER S, JONES S, et al. Okapi at TREC-3[EB/OL]. (2022–01–13)[2024–02–27]. https://api.semanticscholar.org/CorpusID:3946054.

[5] DATTA S, GANGULY D, ROY D, et al. Overview of the causality-driven adhoc information retrieval (CAIR) task at FIRE-2021[C]//Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation. Virtual Event: ACM, 2021: 25–27.

[6] 庞亮, 兰艳艳, 徐君, 等. 深度文本匹配综述 [J]. 计算机学报, 2017, 40(4): 985–1003.
PANG Liang, LAN Yanyan, XU Jun, et al. A survey on deep text matching[J]. Chinese journal of computers, 2017, 40(4): 985–1003.

[7] HUANG Posen, HE Xiaodong, GAO Jianfeng, et al. Learning deep structured semantic models for web search using clickthrough data[C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco: ACM, 2013: 2333–2338.

[8] SHEN Yelong, HE Xiaodong, GAO Jianfeng, et al. Learning semantic representations using convolutional neural networks for web search[C]//Proceedings of the 23rd International Conference on World Wide Web. Seoul: ACM, 2014: 373–374.

[9] PALANGI H, DENG L, SHEN Y, et al. Semantic modelling with long short-trem memory for infomation retrieval[EB/OL]. (2015–05–27)[2022–01–13]. https://arxiv.org/pdf/1412.6629.

[10] GUO Jiafeng, FAN Yixing, AI Qingyao, et al. A deep relevance matching model for ad-hoc retrieval[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. Indianapolis: ACM, 2016: 55–64.

[11] HUI Kai, YATES A, BERBERICH K, et al. PACRR: a position-aware neural IR model for relevance matching [EB/OL]. (2017–04–12)[2024–02–27]. http://arxiv.org/abs/1704.03940.pdf.

[12] HUI Kai, YATES A, BERBERICH K, et al. Co-PACRR: a context-aware neural IR model for ad-hoc retrieval[C]//Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. Marina Del Rey: ACM, 2018: 279–287.

[13] ALTNEL B, GANIZ M C. Semantic text classification: a survey of past and recent advances[J]. Information processing & management, 2018, 54(6): 1129–1153.

[14] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013–01–16)[2024–02–27]. http://arxiv.org/abs/1301.3781.pdf.

[15] LU Yiwei, YANG Ruopeng, JIANG Xuping, et al. Research on military event detection method based on BERT-BiGRU-attention[C]//2021 IEEE International Conference on Consumer Electronics and Computer Engineering. Guangzhou: IEEE, 2021: 1–5.

[16] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018–10–11)[2024–02–27]. http://arxiv.org/abs/1810.04805.pdf.

[17] 于润羽, 杜军平, 薛哲, 等. 面向科技学术会议的命名实体识别研究 [J]. 智能系统学报, 2022, 17(1): 50–58.
YU Runyu, DU Junping, XUE Zhe, et al. Research on named entity recognition for scientific and technological conferences[J]. CAAI transactions on intelligent systems, 2022, 17(1): 50–58.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017: 6000–6010.

[19] JIANG Teng, ZHANG Zehan, YANG Yupu. Modeling coverage with semantic embedding for image caption generation[J]. The visual computer, 2019, 35(11): 1655–1665.

[20] 蔡银琼, 范意兴, 郭嘉丰, 等. 基于多表达的第一阶段语义检索模型 [J]. 计算机工程与应用, 2023, 59(4): 139–146.
CAI Yinqiong, FAN Yixing, GUO Jiafeng, et al. Multi-representation model for the first-stage semantic retrieval[J]. Computer engineering and applications, 2023, 59(4): 139–146.

[21] 巩轶凡, 刘红岩, 何军, 等. 带有覆盖率机制的文本摘要模型研究 [J]. 计算机科学与探索, 2019, 13(2): 205–213.
GONG Yifan, LIU Hongyan, HE Jun, et al. Research on text summarization model with coverage mechanism[J]. Journal of frontiers of computer science and technology, 2019, 13(2): 205–213.

[22] SEE A, LIU P J, MANNING C D. Get to the point: summarization with pointer-generator networks[EB/OL]. (2017–04–14)[2024–02–07]. http://arxiv.org/abs/1704.04368.

[23] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al.

Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. (2014–06–03)[2024–02–07]. http://arxiv.org/abs/1406.1078.pdf.

[24] NGUYEN T, ROSENBERG M, SONG X, et al. A human generated machine reading comprehension dataset [EB/OL]. (2018–10–31)[2024–02–07]. https://arxiv.org/pdf/1611.09268.pdf.

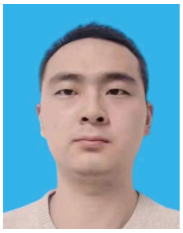
[25] DAI Zhuyun, CALLAN J. Context-aware term weighting for first stage passage retrieval[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 1533–1536.

[26] NOGUEIRA R, YANG Wei, LIN J, et al. Document expansion by query prediction[EB/OL]. (2019–04–07)[2024–02–07]. http://arxiv.org/abs/1904.08375.

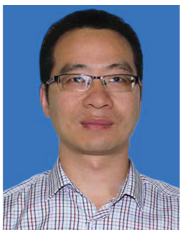
[27] ZHAN Jingtao, MAO Jiaxin, LIU Yiqun, et al. RepBERT: contextualized text embeddings for first-stage retrieval[EB/OL]. (2020–06–28)[2024–02–07]. http://arxiv.org/abs/2006.15498.pdf.

作者简介:

邸剑, 高级工程师, 主要研究方向为人工智能及应用、物联网技术及应用、大数据与云计算。先后主研、参研科技项目 20 余项, 获省部级科技进步奖 2 项, 获授权发明专利 1 项。发表学术论文 30 余篇, 参编教材 1 部。E-mail: dijian6880@163.com。



刘骏华, 硕士研究生, 主要研究方向为深度学习、自然语言处理。E-mail: 220192221061@ncepu.edu.cn。



曹锦纲, 讲师, 博士, 主要研究方向为图像处理和模式识别。发表学术论文 10 余篇。E-mail: caojg168@126.com。

2024 第二届全国人工智能应用场景创新挑战赛

中国人工智能学会、科技部新一代人工智能发展研究中心联合主办的“‘场景驱动·数智强国’——2024 第二届全国人工智能应用场景创新挑战赛”正在火热报名中。大赛采用“开放专题”和“揭榜挂帅”两类竞赛模式, 按照网络选拔赛、省级专项赛和全国总决赛三级赛制进行比拼, 公开征集面向人工智能工程应用、开创性技术突破和产业化落地的创新项目参赛挑战。通过网络选拔、重点推荐、行业晋级、路演比拼、科奖嘉年华等重点环节角逐, 择优推荐总决赛获奖项目进入《全国人工智能应用场景优秀案例目录》。

参赛项目以团队为单位报名参赛。允许跨校和跨企业组建参赛团队(参赛申报单位不超过 4 家, 牵头单位 1 家排首位, 申报组别以牵头单位为准), 参赛团队所报参赛创业项目, 须为本团队经营的项目, 禁止借用冒用他人项目参赛。根据参赛申报人或创新创业团队所处阶段, 项目分为高校种子组、企业天使组、企业成长组并按照五大专题赛道所属应用场景设置参赛项目类型。

参赛报名截止时间 2024 年 7 月 31 日 24:00, 请感兴趣的单位或团队选手登录全国人工智能应用场景创新挑战赛官方网站(www.cicas.cn)查看参赛相关事宜。

联系方式

全国人工智能应用场景创新挑战赛组委会秘书处

联系人: 王老师　15726613955, 杨老师　18994413779

邮　箱: zwh@cicas.cn

网　址: www.cicas.cn